# CUESTIONARIO UD_01

1. ¿Que problema de base origina la aparición de las metodologías y tecnologías Big Data?

- El tener datos que no se sabe de dónde proceden.
- El tener grandes cantidades de datos que no caben en el almacenamiento conjunto de varias máquinas.
- **El tener grandes cantidades de datos que desbordan los recursos de máquinas individuales.**
- La incapacidad de realizar analítica en una única máquina.

2. Si en algún atributo de la especificación de un dispositivo hardware vemos unvalor de 1 kB, ¿a cuántos bytes corresponde?

- A 1000 bytes siempre.
- A 1020 bytes siempre.
- **Dependiendo de la situación, quizás se refiera a 1 kB (que corresponde a1000 bytes), o a 1KiB (que corresponde a 1024 bytes).**
- A 1024 bytes siempre.

3. ¿A qué tipo de información corresponde, generalmente, un fichero con extensión .json?

- Estructurados.
- No estructurados.
- **Semiestructurados.**
- Metadatos
-

4. ¿A qué nos referimos cuando decimos que hay ruido en los datos?

- A que el fichero de audio se grabó con un micrófono de baja calidad.
- A que guardamos el sonido en un ambiente ruidoso.
- **A que parte de los datos no contienen información usable o de la que se pueda obtener algún tipo de valor.**
- No puede haber ruido en los datos.

5. Si los datos van perdiendo valor con el tiempo y tenemos muchos datos antíguos, ¿merece la pena utilizarlos siempre junto con los más nuevos?
   **Depende de la situación en que se vayan a utilizar**

6. ¿Cuáles de los siguientes son posibles beneficios de las metodologías y tecnologías Big Data?

- **Soportar la toma de decisiones.**
- **Mejorar las operaciones en empresas e instituciones.**
- **Ayudar a detectar enfermedades.**
- **Ayudar a los científicos a realizar nuevos descubrimientos.**

7. ¿Cuáles de los siguientes son eventos susceptibles de generar datos?
- **Un pago con tarjeta.**
- **Un alta de usuario en una web.**
- **Una medida de presión atmosférica en una estación meteorológica.**
- **Un análisis de sangre de un paciente en un hospital.**

8. ¿Qué hacemos si un clúster necesita más capacidad de almacenamiento?
- Hacemos escalado vertical de todos los nodos, aumentando el tamaño de almacenamiento de cada uno.
- **Hacemos escalado horizontal, añadiendo mas nodos al clúster.**
- Hacemos escalado en diagonal, aumentando el almacenamiento en los nodos que tengan menos espacio disponible.
- Hacemos escalado vertical, añadiendo mas nodos al clúster

9. Las Bases de Datos Relacionales ofrecen un alto rendimiento para realizar transacciones, pero sus motores no están pensados para el caso en el que una tabla sea tan grande que todos sus registros no puedan ser almacenados dentro de un mismo servidor. De modo que, ¿son las Bases de Datos Relacionales apropiadas para entornos Big Data? **No, las bases de datos relacionales son las mayores enemigas del Big Data** debido a limitaciones en escalabilidad.

10. ¿Cuál de las siguientes afirmaciones es cierta en relación a las bases de datos relacionales?
- Utilizan MySQL como lenguaje de consulta.
- No es necesario conocer los tipos de datos que se van a almacenar desde un primer momento, sino que se determina al realizar su lectura.
- Podemos utilizar el tipo de datos RDBMS, en el cual cabe cualquier número de bytes ya que se guarda en ficheros específicos fuera de la base de datos.
- **Si creamos índices para las columnas sobre las que vayamos a hacer búsquedas, éstas se ejecutarán más rápido.**

11. ¿Cuál de las siguientes afirmaciones es correcta respecto de un dataset?
- Siempre vienen en ficheros de texto plano.
- Contienen datos de usuarios.
- No pueden contener datos de usuarios porque constituye un uso prohibido.
- **Que contenga imágenes no significa que no pueda contener también texto, audio o vídeo**.

12. ¿Un almacén de datos puede incluir en su interior una base de datos relacional?
- **Sí.**
- No, sólo puede incluír subsistemas OLAP.
- No, sólo bases de datos de tipo NoSQL.

13. Las bases de datos relacionales se usan en el día a día para operaciones transaccionales. ¿Son, por lo tanto ACID? **Si, ya que deben cumplir que vayan a ser usadas para realizar transacciones**

14. Aunque los cortes de comunicación entre nodos son poco frecuentes, lo cierto es que pueden ocurrir en cualquier momento, y por lo general ninguna institución ni empresa está dispuesta a que su base de datos distribuída deje de funcionar durante esos momentos. En esa gran cantidad de casos en los que se quiere cumplir con la tolerancia a particionamiento, ¿qué opciones tenemos?
**Nunca puede cumplirse C+A+P, sino que habrá que escoger siempre entre C+A, C+P o A+P a la hora de diseñar la base de datos distribuida.**

15. Dado que las bases de datos distribuídas que emplean la filosofía BASE dan prioridad a la disponibilidad a costa de la consistencia, ¿son una buena elección para uso transaccional? **No son la mejor opción dado que las transacciones tradicionales suelen basarse en el principio ACID (Atomicidad, Consistencia, Aislamiento/Isolate y Durabilidad), y al no cumplir estos requisitos puede ser un problema de coherencia.**

16. Pero entonces, si sólo tenemos un único procesador que no es multihilo y el gestor de procesos del sistema operativo reparte tiempos entre los procesos, ¿podemos decir que está realizando varias tareas al mismo tiempo? **No, ya que el procesador ejecuta una tarea a la vez, alternando entre diferentes procesos según el esquema de planificación del sistema operativo.**

17. ¿Cuál de las siguientes afirmaciones es correcta en relación a paralelización de tareas?
- Todas las tareas pueden paralelizarse de modo que se ejecuten más rápido que sin paralelizar.
- **No todas las tareas pueden paralelizarse.**
- El mayor problema de la paralelización es el tiempo que se tarda en integrar los datos resultantes de tratar cada fragmento.

18. A la hora de enviar datos entre dos nodos de un clúster, ¿en qué caso será más rapida la comunicación?
- Es vital que haya pocos metros de cable entre los nodos.
- **Será más rápido cuantos menos saltos entre switches haya que hacer para llegar de un nodo a otro.**
- Lo más rápido siempre es enviarlos a otro CPD si éste cuenta con máquinas más rápidas.
- Los nodos de un clúster no pueden comunicarse entre sí.

19. ¿Cuál es la diferencia entre procesamiento en tiempo real y procesamiento en streaming?

- **Tiempo real implica que los resultados se producen en poco tiempo, mientras que en streaming implica que es capaz de tener en cuenta datos que van entrando constantemente.**
- Son lo mismo.
- Streaming significa que es transaccional, mientras que en tiempo real significa que no es transaccional.
- Streaming significa que el procesamiento es rápido, mientras que en tiempo real significa que todo ocurre a la velocidad a la que lo solicita el usuario.

20. Cuando hablamos de OLTP, ¿qué tipo de base de datos se está empleando por lo general?
- Una base de datos orientada a grafos.
- No será una base de datos común sino que todo se estará almacenando en la memoria RAM del sistema.
- **Una base de datos relacional.**
- No se emplea una base de datos sino un almacenamiento distribuído como HDFS o S3.

21. ¿Cuáles son rasgos típicos de las estructuras de datos empleadas para OLAP?
- Almacenamiento en unidades SSD para un acceso más rápido.
- Almacenamiento en memoria RAM de estructuras multidimensionales.
- Almacenamiento en memoria RAM de datos previamente normalizados.
- **Estructuras multidimensionales que se almacenan en sistemas distribuídos tipo HDFS o S3.**

22. Dado que en ambientes de Big Data el ser capaz de manejar grandes volúmenes de datos (V) es una obligación, ¿podremos típicamente realizar analítica en tiempo real con todos ellos? **Típicamente no es práctico ni viable, dado que requiere sistemas rápidos y eficientes para recibir datos en tiempo real.**

23. Dado que en ambientes de Big Data el ser capaz de manejar grandes volúmenes de datos (V) es una obligación, ¿podremos típicamente realizar procesamiento por lotes empleando todo ese conjunto de datos? **Todos los datos no se procesan por lotes (Batch) debido a la ineficiencia y tiempo que puede requerir. Aunque es posible realizarlo, es más eficiente y práctico hacerlo en partes o dependiendo de las necesidades específicas**

24. ¿Qué capa se encarga de integrar los datos de modo que queden unificados con sentido propio para la tarea que se va a realizar con ellos?
- La capa de ingestión.
- La capa de colección.
- La capa de almacenamiento.
- **La capa de procesamiento.**

25. ¿Qué queda representado en el paisaje de Big Data?

- **Las distintas capas por las que pasan los datos.**
- La posible distribución de los nodos de un clúster dentro de un centro de datos.
- Las herramientas y utilidades que se pueden utilizar.
- Las herramientas y utilidades que sirven para obtener datos de diversas fuentes.